DOCUMENT RESUME

ED 397 117 TM 025 251

AUTHOR Schumacker, Randall E.
TITLE Many-Facet Rasch Model Selection Criteria: Examining Residuals and More.
PUB DATE Apr 96
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Ability; Adults; *Chi Square; *Criteria; Earth Science; Geography; Goodness of Fit; History; Item Response Theory; *Judges; Scores; *Selection
IDENTIFIERS Facet Analysis; FACETS Model; *Rasch Model; *Residuals (Statistics)

ABSTRACT

This research examined the significance of facet selection in a multi-facet Rasch model analysis. The residuals or remaining error in a multi-facet Rasch model were further studied in the context of a full and reduced data-to-model fit chi-square, given the specific design. In addition, main effect facet contributions to person measures and the interaction among elements of two facets were investigated. Seventy-four subjects participated, with the variables or facets studied being subjects, judges, sessions, topics, and tasks. Each subject was rated by a sample of 6 of the total of 31 judges on recall, interpretation, and application of history, geography, and earth science domains. Fixed chi-square values were significant for all facets included in the model, indicating that the elements for each facet differed significantly and had different effects on the subject's scores that needed to be accounted for through adjustment to scores or ability estimates. Examination of models in which one facet was excluded further indicated a facet's contribution to the overall data-model fit. The chi-square test can indicate how the facet elements differ, and calibrated measures indicate how much the subject ability estimates should be adjusted to account for the characteristics of the particular elements encountered by a subject. Appendix A shows entry of the original coded data, and Appendix B presents sample measurement report. (Contains one figure, five tables, and five references.) (SLD)

MANY-FACET RASCH MODEL SELECTION CRITERIA: EXAMINING RESIDUALS AND MORE

Randall E. Schumacker
University of North Texas

American Educational Research Association
April, 11, 1996
New York, N.Y.

MANY-FACET RASCH MODEL SELECTION CRITERIA:
EXAMINING RESIDUALS AND MORE

The many-facet Rasch model has distinct advantages over generalizability theory (Linacre, 1993; Stahl & Lunz, 1993). These advantages include the use of person measures rather than raw scores and the adjustment of person measures for facets included in the model. Another advantage of the many-facet Rasch model is the facet "connectedness" that is required if linear "rulers" are to be created for each facet. The use of crossed and nested designs is similar in both methods, however, in a many-facet Rasch model adjustments for facet "effects" can be undertaken, even in the presence of missing data. In the generalizability theory analyses, no adjustments are made, rather a G-study estimates the variance components and a D-study is used to determine if a sufficient number of conditions (i.e., items, raters, occasions) exist to yield dependable scores. This approach could obviously require that subjects be measured again based upon the D-study results or another sample be measured. In many instances this is simply not possible nor prudent, e.g., professional licensure. There is also no guarantee that the G-study variance component estimates will be the same in a subsequent sample or when re-testing subjects that another D-study wouldn't indicate different conditions being required. The generalizability analyses are further limited in that they use raw scores and require complete data.

From a design perspective, raw scores are obtained from individuals under certain defined facets, which in the many-facet Rasch model, are converted to linear logit measures. The particular facet elements a candidate encounters effects the subject's scores making it important to determine whether the facet elements are significantly different. The extent to which facet elements effect individual scores is found by examining the facet calibrations and noting differences among facet elements. Additionally, one should be able to determine if an interaction between the elements of two facets influence a persons' score. It is therefore important to be able to test both main effects and interaction effects in many-facet Rasch models.

This research examined the significance of facet selection in a multi-facet Rasch model analysis. The residuals or remaining error in a multi-facet Rasch model was further studied in the context of a full and reduced data-to-model fit X^2, given the specific design. In addition, main effect facet contributions to person measures and the interaction among elements of two facets was investigated.

METHODOLOGY

Data

A total of seventy-four (74) subjects participated in the study. The variables or facets studied included subjects, judges, sessions, topics, and tasks. The session facet was coded from 1 to 5 and represented the day of the week in which each subject was rated by a sample of six judges. The topic facet included three elements: history, geography, and earth science. The task facet included three elements: recall, interpretation, application. A total of thirty-one (31) judges rated subjects on the tasks within each topic, however, the same judges did not rate all subjects. There were no missing data.

Two different judges provided a rating of 0 = 'F', 1= 'D', 2= 'C', 3= 'B', 4= 'A' on the three tasks within each of the three topics for a given subject. Consequently, each subject received a total of eighteen (18) ratings based on the tasks within the topics. Raw scores could range from 0 to 72. The sample mean = 52, standard deviation = 9, and the sample size = 74. The facet's analysis will compute a calibrated person measure which accounts for the particular facet elements encountered by the subject. The conversion of these calibrated person measures or logits to percents or probability of correct response can be found in Wright and Stone (1979, p. 36).

Design

Each subject was rated on three tasks in one of three topics by two judges on only one of five different days. The research design indicating ratings for one subject only is depicted in Figure 1. This design would be characterized as

a nested design because subjects were nested within each session and not all judges rated all subjects. It also contained a crossed design effect with elements of the task facet crossed with elements of the topic facet. This design must therefore consider the influence that each of these facets might have upon subject scores or measures. From a design perspective, it was determined that, the session a subject was rated in, the two judges and topic, as well as, the task difficulty, would impact upon a subject's score. This resulted in a five-facet Rasch model which included main effects for subject, session, judge, topic and task facets.

Insert Figure 1 Here

Analysis

The basic Rasch model using dichotomous scoring (1=correct, 0=wrong) is depicted as: $\log [P_{ni1} / P_{ni0}] = B_n - D_i$. Where

P_{ni1} = probability of subject n getting item i correct (x=1)
P_{ni0} = probability of subject n getting item i wrong (x=0)
B_n = ability of student n
D_i = difficulty of item i.

This Rasch model has two facets, subject ability and item difficulty. In multi-facet Rasch models, this basic model is expanded to include other facets. The five-facet Rasch model used for the analysis is written as follows: $\log [P_{nijmsx} / P_{nijmsx-1}] = B_n - D_i - C_j - T_m - S_s - F_{ix}$. The terms are defined as follows:

P_{nijmsx} =probability of student n being rated x on task i in topic m in session s by judge j.

$P_{nijmsx-1}$=probability of student n being rated x-1 on task i in topic m in session s by judge j.

B_n =ability of student n

D_i =difficulty of task i

C_j =effect of rating by judge j

T_m =effect of rating for topic m

S_s =effect of rating in session s

F_{ix} =difficulty of rating step x relative to step x-1 (rating scale categories).

The raw scores are input and the data re-written for suitable input into a FACET program using a Facform program (Linacre, 1994; see Appendix A). The FACET program outputs several different types of chi-square tests. These chi-square tests are called "fixed" effects, "random" (normal), and data-to-model "residual" fit. An understanding of each is important in making decisions about facet inclusion in a model, facet level similarity, and facet level interaction. To examine the similarity among facet elements a "fixed" effects chi-square test is possible which can test whether the L measures are statistically equivalent to one common "fixed" effect apart from measurement error. The basic formula is: $X^2 = \Sigma (w_i D^2_i) - (\Sigma w_i D_i)^2 / \Sigma w_i$ with L-1 degrees of freedom. The value w_i, computed as $1/SE^2_i$, indicates the information for L measures, D_i, with standard errors, SE_i. If $p > .05$, then L facet element measures are statistically equivalent. For example, the "fixed" effects chi-square test for testing the similarity of judges would reflect a test of the following null hypothesis: H_o: $judge_i = judge_j$, where $i \neq j$.

A "random" (normal) chi-square test is also possible for each facet included in a model. The formula is: $X^2 = \Sigma z^2_L$ where L represents the elements of the facet. The "random" chi-square has an expected value equal to the number of elements in the facet. For example, topic had three elements so the expected "random" (normal) chi-square value is 3. There were 31 judges so the expected "random" (normal) chi-square value is 31, and so forth for the other

facets. The "random" (normal) chi-square is a general homogeneity test of logit measure distribution normality.

A data-to-model "residual" fit chi-square test is also possible where the sum of squared standardized residuals equals a chi-square value with degrees of freedom equal to the number of measurable responses minus the number of independent estimable parameters. The number of measurable responses in the present example is 1,332 (74 subjects times 18 ratings). The data-to-model "residual" fit X^2 is the sum of the squared standardized residuals for these 1,332 measures. The "residual" fit chi-square is useful for testing the effect of including a facet in the model. A full model can be run and the "residual" fit chi-square noted. Then a reduced model excluding one facet can be run, once again noting the "residual" fit chi-square value. A difference between the two "residual" fit chi-square values with $L_2 - L_1$ degrees of freedom would then indicate the unique "effect" or contribution of the facet. Basically, if a facet doesn't have an "effect" in the model, then there should be little difference in the standardized residual values.

RESULTS

Facet Main Effect Contributions

The five-facet Rasch model included subjects, topics, tasks, judges, and session effects. These facets were selected based on how subject's scores were obtained and what "conditions" might affect them. Obviously, the creation of a proper research design is instrumental to interpreting and understanding results (Lunz, 1994). The "fixed" chi-square value for each facet is presented in Table 1. The "fixed" chi-square values were significant for all facets included in the model. This indicates that the elements for each facet, differed significantly. The interpretation of differences among facet elements relates to how the subjects' scores are affected by the particular combination of facet elements encountered. Subject ability estimates are adjusted according to the location of

the elements encountered on the scale. Significant differences in facet elements indicates the need for this adjustment. The facet element measures for judges, sessions, topics, and tasks are presented in Appendix B.

Insert Table 1 Here

Residual analysis (full and reduced models)

Whether a given facet contributes to a model, above and beyond, the contribution of other facets in the model can be tested as a "residual" chi-square difference test. Table 2 shows the "residual" chi-square values for the five-facet model and four-facet models in which a different facet was dropped. The chi-square difference test indicates whether a particular facet contributes significantly to the model. The full model was significant at the .01 level of significance. In the four-facet reduced models, dropping the judge facet resulted in a non-significant model at the .01 level of significance. The judge facet significantly affected the measurement model, and therefore, the subject's scores. An inspection of the residual values for each subject from the five- and four-facet models would further reveal the influence that different judges had upon subject's scores.

Insert Table 2 Here

Facet Interaction

In addition to testing for main effects of facet inclusion or differences between elements of a facet, interaction effects between the elements of two facets can be investigated. In the present example, an interaction between elements of the session and judge facets were hypothesized in the full model.

This was deemed important because different judges rated subjects nested in a given session (day of week). Consequently, the judge's ratings could differ by session thereby affecting subjects' scores. An examination of interaction is possible in the FACETS program using the individual "bias" estimates and associated z-scores (measure divided by standard error) for each combination of judge and session element. If the z-score associated with a "bias" measure is greater than ± 2, a significant difference exists between the "observed" score and "expected" score. A significant difference between these two scores indicates that the judge rated subjects in that session significantly different than expected based on his/her performance in all sessions.

Insert Table 3 Here

Table 3 is in an abbreviated form and only presents the interacton bias measures for selected judges with the elements of the session facet. To illustrate, subject number 102 had a raw score of 53, a calibrated person measure of 2.42, and was rated by judges 1, 12, 16, 29, 33, and 34 in session one. The interaction effect is determined by adding the bias measures for these judges in session one which yields .86 (.47 + .94 + -.15 + -.67 + -.16 + .43). For comparison purposes, subject 315 had a raw score of 53, a calibrated person measure of 1.46, and was rated by judges 8,12,14,26,31,and 33 in session three. The interaction effect upon this subject's score was -.20 (-.24 + -1.00 + -.77 + -.29 + 1.25 + .85). This indicates, that although these two subjects had identical raw scores, each was affected by which judges, in which sessions, rated them; hence the different calibrated person measures.

The "fixed" chi-square value in the summary table can also be partitioned into the elements of the facet for judge and/or session. The partitioned chi-square values should sum to the global (total) "fixed" chi-square value. Moreover, differences in these chi-square values can yield "simple" effects tests between levels of a facet. Table 4 presents the "fixed" chi-square values for the five sessions. Judge ratings in the second and fifth sessions were different

from ratings given in the third session. Table 5 presents the "fixed" chi-square values for the thirty-one judges. Judges 704, 706, 710, 713, 720, and 728 rated more severely overall and had chi-square values significantly higher than the expected value of 5. A wide variation in ratings by a judge, either severe and/or lenient across sessions will increase the chi-square value. For example, judge 710 rated the sessions as follows: (1) 2.50; (2) -.32; (3) -1.56; (4) -3.00 (5) 2.52. Judge 710 was therefore rating severely in sessions one and five, but lenient in sessions three and four. Overall, this variation indicates an inconsistent judge.

If $X^2 = 0$, then no difference exists between the observed scores and the expected scores obtained from a judge's ratings. If X^2 values are between 0 and 5, then a judge has given lenient ratings, i.e., observed scores are greater than expected scores. If $X^2 > 5$, then expected scores are higher than observed scores indicating more severe ratings. The range of z-scores, however, must be taken into consideration. For example, judge 726 would be considered a lenient rater, however, only small differences exist between observed and expected scores across the sessions indicated by the narrow range of z-score values (-.47 to .82). Judge 726 is therefore consistent in rating subjects across sessions and has only small differences between observed and expected scores. This is where an examination of each judge's z-score, infit, and outfit for each session can detect variations in ratings across sessions.

Insert Tables 4 & 5 Here

CONCLUSIONS

The five-facet model was hypothesized based upon a test design that required subject, session, judge, topic, and task facets. The main effects for each facet was examined for significance using a "fixed" X^2 value. All facets had a significant "fixed" X^2 value indicating that the elements of each facet were significantly different. These differences are apparent upon inspection of

the facet measures in the measurement reports in Appendix B. Consequently, the elements of the facets have different effects upon the subject's scores and need to be accounted for through adjustment to their scores or ability estimates.

A further examination of the reduced four-facet models, in which one and only one facet was excluded, can further indicate a facets contribution to the overall data-to-model fit. If a facet is removed, and the residual X^2 value of the model becomes non-significant, the facet significantly affects scores. The X^2 difference test between the full model and each reduced four-facet model indicated that only the judge facet significantly reduced the model fit.

An examination of interaction between elements for session and judge was hypothesized based upon the design. The z-score values for judges in the sessions best indicates the effect of variation or consistency of ratings upon subject's scores. Simple effects tests in which the global "fixed" X^2 value is partitioned into the levels of the facet were possible. Judges were not always in agreement across the sessions. This could be because the judges were different or because the candidates were systematically more or less able within sessions. The nested design makes it impossible to determine the reason for observed differences across sessions. Also, simple effects for the judges themselves revealed that some judges were more severe than others.

This paper discussed a X^2 test of facet main effects, a X^2 test of differences between levels of a facet, a X^2 difference test for facet contribution to the model, a method for examining interaction (z-scores), and a partitioning of the global chi-square value into a "simple" effects X^2 test. These X^2 tests were presented in the context of a test design. If the elements of a facet are significantly different, then the facet elements encountered by a subject should be accounted for when computing a subject's ability estimate. After all, the primary intent of the many-facet Rasch model is not to maximize the data-to-model fit, rather to construct generalizable linear measures for subjects, including standard error (reliability) and fit (validity).

From a design perspective to the inclusion of facets, the intent is to reduce measurement error and correctly estimate person measures (Lunz, 1994).

Being able to test whether a facet has a significant effect upon subject's scores permits attention to properly adjusting scores. An examination of the calibrated measures for each element of a facet indicates the particular amount of adjustment to be made to the person ability estimates. A X^2 test indicates that the facet elements differ, the calibrated measures indicate how much the subject ability estimates should be adjusted to account for the characteristics of the particular elements encountered by a subject. Herein lies the specificity we seek.

REFERENCES

Linacre, M. (1993, April). *Generalizability theory and many-facet Rasch measurement*. Paper presented at the annual meeting of the American Educational Research Association. Atlanta, Georgia.

Linacre, M.J. (1994). *A User's Guide to Facets*.
MESA Press: Chicago, IL.

Lunz, M.E. (1994, October). *Reducing the error of measurement by design for performance examinations*. Paper presented at the annual meeting of the Mid-Western Educational Research Association. Chicago, Illinois.

Stahl, J.A. & Lunz, M.E. (1993, April). *A comparison of Generalizability theory and Multi-faceted Rasch measurement*. Paper presented at the annual meeting of the American Educational Research Association. Atlanta, Georgia.

Wright, B.D. & Stone, M.H. (1979). *Best Test Design*.
MESA Press: Chicago, IL.

Table 1. Facet Main Effects (full model)

Facet	Fixed X^2	df	p
Subjects	877.3	73	<.01
Session	12.8	4	.01
Topic	8.6	2	.01
Judge	223.9	30	<.01
Task	26.8	2	<.01

Note: Data-to-model global fit residual chi-square = 1345, df = 1217, p < .01; df = 1332 minus 115 estimable parameters.

Table 2. Full model (5-facets) compared to reduced models (4-facets)

Facet	Residual X^2	df	p	estimable parameters
Full model	1345.0	1217	.01	115
Reduced models:				
No Session	1312.1	1221	.03	111
No Judge	1315.2	1247	.09	85
No Topic	1320.5	1219	.02	113
No Task	1321.4	1219	.02	113

Note: Full model included facets for subject. session, judge, topic, and task.

Table 3. Interaction bias measures for judge and session facet

Obsvd Score	Exp. Score	Obsvd Count	Obs-Exp Average	Bias Measure	Model S.E.	Z-Score	Infit MnSq	Outfit MnSq	session	measure	judge	measure
17	17.8	6	-.14	[illegible]	.70	.60	0.4	0.4	First	.10	01	.27
21	22.7	9	-.19	.42	.48	.88	0.4	0.4	Second	.03	01	.27
22	22.2	9	-.02	.05	.49	.09	0.8	0.9	Third	-.08	01	.27
26	26.6	9	-.06	.18	.56	.32	0.7	0.6	Fourth	.21	01	.27
33	29.5	9	.39	-1.43	.71	-2.01	1.1	1.2	Fifth	-.26	01	.27
23	23.8	9	-.09	.20	.50	.40	0.9	1.0	First	.10	08	.57
20	20.0	9	.01	-.01	.46	-.02	1.3	1.4	Second	.03	08	.57
27	26.3	9	.08	-.23	.58	-.40	0.6	0.6	Third	-.08	08	.57
21	23.8	9	-.31	.65	.46	1.42	1.6	1.5	Fourth	.21	08	.57
27	24.1	9	.33	-.89	.57	-1.56	0.4	0.4	Fifth	-.26	08	.57
23	26.0	9	- 33	.90	.52	1.71	0.7	0.6	First	.10	12	.25
27	28.4	9	-.15	.47	.58	.81	1.7	1.7	Second	.03	12	.25
30	27.2	9	.31	-.99	.61	-1.62	1.2	1.1	Third	-.08	12	.25
21	22.1	9	-.12	.27	.49	.56	1.1	1.0	Fourth	.21	12	.25
29	26.2	9	.31	-.92	.59	-1.54	1.0	1.0	Fifth	-.26	12	.25
29	27.6	9	.15	-.49	.60	-.82	1.2	1.2	First	.10	14	-1.15
25	26.9	9	-.21	.58	.53	1.09	0.7	0.7	Second	.03	14	-1.15
33	31.1	9	.21	-.81	.70	-1.16	1.7	1.4	Third	-.08	14	-1.15
24	25.0	9	-.12	.32	.54	.59	0.5	0.5	Fourth	.21	14	-1.15
31	31.0	9	.00	-.02	.61	-.03	0.7	0.7	Fifth	-.26	14	-1.15
18	17.6	6	.06	-.18	.72	-.26	0.0	0.0	First	.10	16	.50
24	22.0	9	.23	-.55	.54	-1.02	0.2	0.2	Second	.03	16	.50
21	25.3	9	-.48	1.13	.48	2.35	0.2	0.2	Third	-.08	16	.50
24	23.1	9	.10	-.25	.52	-.47	2.0	2.3	Fourth	.21	16	.50
30	28.9	9	.13	-.41	.60	-.67	0.6	0.6	Fifth	-.26	16	.50
26	27.0	9	-.11	.33	.55	.59	0.4	0.4	First	.10	26	-.43
29	30.3	9	-.14	.46	.60	.77	0.5	0.5	Second	.03	26	-.43
30	29.2	9	.09	-.30	.61	-.49	0.4	0.4	Third	-.08	26	-.43
27	26.3	9	.07	-.22	.58	-.38	0.4	0.5	Fourth	.21	26	-.43
30	29.0	9	.11	-.37	.60	-.61	0.5	0.5	Fifth	-.26	26	-.43
18	16.7	6	.22	-.65	.72	-.91	0.0	0.0	First	.10	29	.84
22	21.7	9	.04	-.08	.50	-.17	0.5	0.5	Second	.03	29	.84
25	28.0	9	-.34	.95	.54	1.75	0.1	0.1	Third	-.08	29	.84
20	19.6	9	.04	-.09	.47	-.18	0.3	0.3	Fourth	.21	29	.84
24	23.1	9	.10	-.25	.53	-.47	0.5	0.5	Fifth	-.26	29	.84
29	27.6	9	.15	-.50	.61	-.81	1.0	1.1	First	.10	31	-.56
27	27.9	9	-.10	.30	.58	.52	1.8	1.7	Second	.03	31	-.56
26	29.5	9	-.39	1.21	.57	2.13	0.2	0.2	Third	-.08	31	-.56
28	25.0	9	.34	-.98	.59	-1.67	0.5	0.5	Fourth	.21	31	-.56
29	28.7	9	.03	-.09	.60	-.15	1.1	1.1	Fifth	-.26	31	-.56
19	18.6	6	.07	-.21	.73	-.29	0.3	0.3	First	.10	33	-.14
25	24.3	9	.08	-.21	.56	-.38	0.5	0.6	Second	.03	33	-.14
26	28.4	9	-.26	.81	.57	1.41	0.2	0.2	Third	-.08	33	-.14
31	27.9	9	.34	-1.31	.70	-1.88	0.7	0.5	Fourth	.21	33	-.14
26	27.5	9	-.17	.50	.56	.88	0.5	0.5	Fifth	-.26	33	-.14
15	16.0	6	-.17	.42	.63	.67	1.3	1.3	First	.10	34	1.28
22	25.3	9	-.37	.93	.50	1.85	0.5	0.6	Second	.03	34	1.28
28	26.5	9	.16	-.47	.57	-.81	0.7	0.7	Third	-.08	34	1.28
22	20.7	9	.15	-.29	.47	-.61	1.6	1.5	Fourth	.21	34	1.28
29	27.6	9	.15	-.49	.60	-.82	0.7	0.7	Fifth	-.26	34	1.28

Note: The 31 judges were not numbered consecutively.

Table 4. Chi-square values for levels of session facet

Session	X^2	n	Z_{min}	Z_{max}
one	55.42	30	.02	2.84
two	40.11	30	.00	2.60
three	61.67	31	.10	4.50
four	53.12	30	.02	3.34
five	46.68	30	.00	2.52

Note: Global "fixed" ΣX^2 = 257, df = 151, $p < .001$. Chi-square = 43.77, df=30, $p = .05$ level of signifcance.

Table 5. Chi-square values for judges

Judge	X^2	n	Z_{min}	Z_{max}
01	5.40	5	.11	-2.01
03	5.92	4	-1.31	1.88
04	29.41	4	-2.45	4.50
06	20.09	5	-2.49	3.34
07	9.25	5	-.95	2.60
08	4.87	5	-1.53	1.48
09	3.55	5	-1.06	1.05
10	24.17	5	-3.00	2.52
11	8.62	5	-1.30	2.40
12	9.43	5	-1.64	1.77
13	20.86	5	-2.44	2.84
14	3.17	5	-1.10	1.14
15	2.81	5	-1.00	1.17
16	7.79	5	-1.05	2.40
18	1.96	5	-1.14	.68
19	3.76	3	-1.34	1.40
20	14.45	5	-2.79	1.79
21	5.62	5	-1.65	1.14
22	5.84	5	-1.38	1.46
23	3.51	5	-1.26	.70
24	3.60	5	-1.19	1.33
25	4.85	5	-1.20	1.22
26	1.88	5	-.47	.82
27	7.31	5	-1.07	1.93
28	12.65	5	-2.49	1.93
29	4.36	5	-.92	1.79
30	6.35	5	-.90	2.00
31	8.25	5	-1.62	2.17
32	5.02	5	-1.28	1.22
33	6.70	5	-1.89	1.46
34	5.55	5	-.84	1.83

Note: $\Sigma X^2 = 257.00$ df=151.

Chi-square=11.07, df=5, p=.05 level.
Chi-square= 9.49, df=4, p=.05 level.
Chi-square= 7.82, df=3, p=.05 level.

Figure 1. Research Design indicating ratings for one subject

Session	1			2			3			4			5		
Topic	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c
Task	123	123	123	123	123	123	123	123	123	123	123	123	123	123	123
Judges															
01															
02															
03															
04															
05															
06															
07	323														
08		322													
09															
10															
11															
12															
13															
14															
15			222												
16															
17															
18															
19															
20															
21	322														
22		333													
23															
24															
25															
26															
27															
28															
29															
30															
31			233												

APPENDIX A

The original coded data was entered as follows:

ID	Session	Subject	Judge1	T11	T21	T31	Judge2	T21	T22	T23
1	1	1	1	3	4	3	2	3	4	3
1	1	2	3	4	3	2	4	2	3	3
1	1	3	5	4	4	4	6	3	3	3
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
74	5	1	11	3	3	3	9	4	3	2
74	5	2	8	2	2	3	1	3	4	4
74	5	3	31	2	2	2	10	3	3	3

The first three lines of coded data indicates one subject (ID variable) who has been rated during session one (Monday) in three subject areas. For each subject area, two different judges have provided ratings on the three tasks for a total of eighteen ratings. The last three lines indicate the last subject rated during session five (Friday) in three subject areas.

The Rasch Facform program converts this data set into six (6) lines per subject with comma separated variables. The raw data for the first subject would be recoded as follows:

```
1,1,1,1,1-3,3,4,3
1,1,1,2,1-3,3,4,3
1,1,2,3,1-3,4,3,2
1,1,2,4,1-3,2,3,3
1,1,3,5,1-3,4,4,4
1,1,3,6,1-3,3,3,3
```

The values between each comma, respectively, are: subject, session, topic, judge, levels of task facet, i.e., 1 to 3, task rating one, task rating two, and task rating three. The total number of data lines in the Rasch facform data file is n = 444 (74 subjects x 6 records).

APPENDIX B

Session Measurement Report

Session	n	Measure	S.E.	Subjects
Monday	252	.10	.11	1-14
Tuesday	270	.03	.10	15-29
Wednesday	270	-.08	.10	30-44
Thursday	270	.22	.10	45-59
Friday	270	-.27	.11	60-74
	1332	.00	.10	

Note: Fixed X^2 = 12.8, df=4, p = .01

Task Measurement Report

Task	n	Measure	S.E.
Recall	444	-.32	.08
Interpretation	444	.25	.08
Application	444	.07	.08
	1332	.00	.08

Note: Fixed X^2 = 26.8, df=2, p<.01

Topic Measurement Report

Topic	n	Measure	S.E.
History	444	-.09	.08
Geography	444	.19	.08
Earth Science	444	-.10	.08
	1332	.00	.08

Note: Fixed X^2 = 8.6, df=2, p=.01

Judge Measurement Report

Judge	n of ratings	Measure	S.E.
01	42	.27	.26
03	33	1.33	.26
04	30	-1.21	.39
06	48	-.61	.25
07	45	-.50	.26
08	45	.59	.23
09	45	.85	.24
10	45	.24	.25
11	45	.59	.23
12	45	.27	.25
13	45	-.35	.26
14	45	-1.20	.27
15	45	.25	.24
16	42	.51	.25
18	45	.89	.22
19	27	-.94	.34
20	45	.00	.25
21	45	-.23	.25
22	42	-.90	.29
23	45	-.29	.26
24	45	-.81	.26
25	48	.48	.22
26	45	-.44	.27
27	45	.90	.23
28	42	-.17	.26
29	42	.90	.24
30	42	-.56	.27
31	45	-.58	.26
32	45	-.47	.26
33	42	-.16	.27
34	42	1.36	.25
	1332	.00	.26

Note: Fixed X^2 =223.9, df=30, $p < .01$